Inversion Medians Outperform Breakpoint Medians in Phylogeny Reconstruction from Gene-Order Data

Bernard M.E. Moret

Department of Computer Science
University of New Mexico
Albuquerque, NM 87131
moret@cs.unm.edu

Coauthors and Sponsors

Coauthors:

- Adam C. Siepel

  Computer Science and Engineering, University of California at Santa Cruz
- Jijun Tang *and* Tao Liu

  Computer Science, University of New Mexico

Sponsor:

- National Science Foundation

  grants EIA 01-21377, EIA 01-13095, and DEB 01-20709

Overview

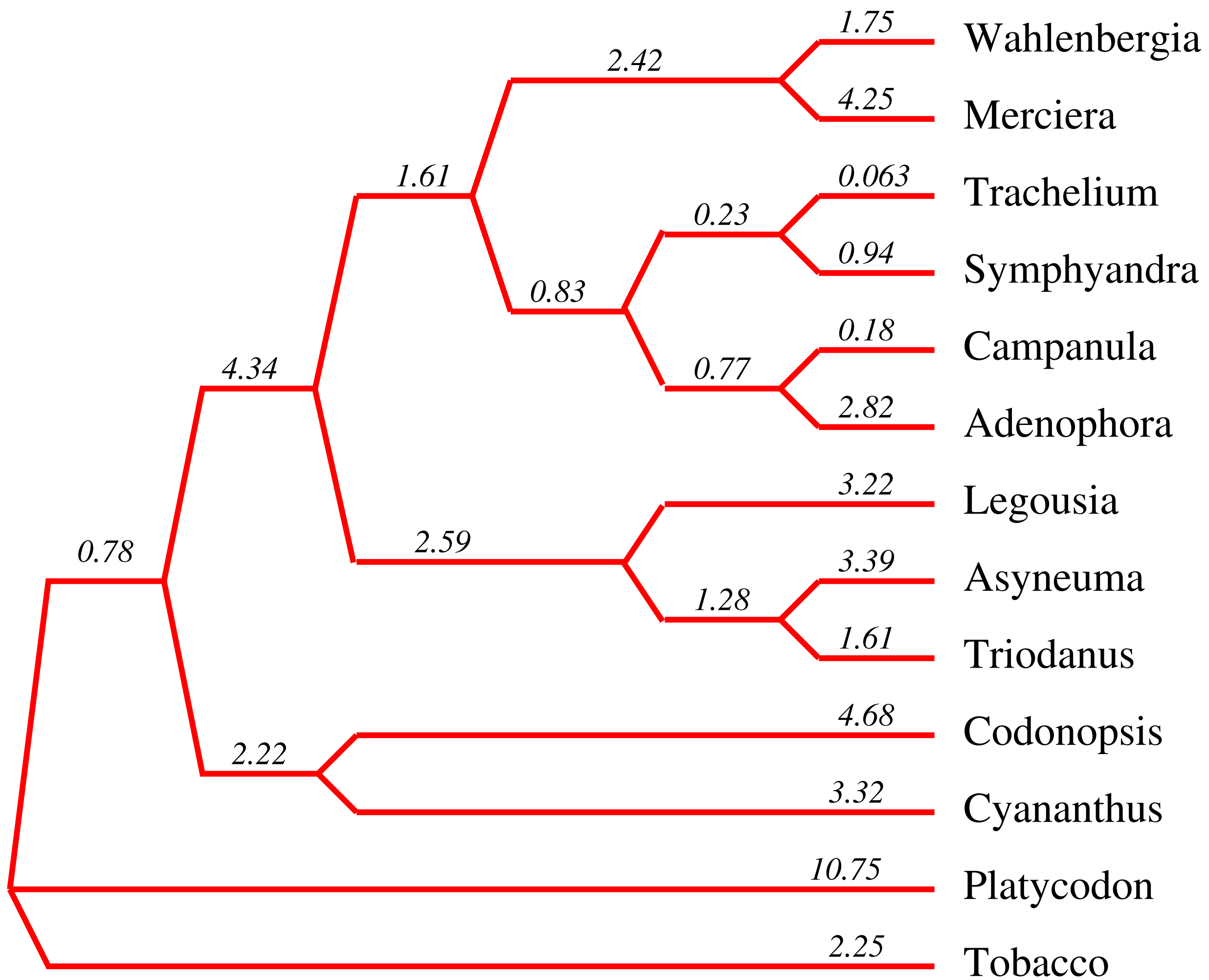
- Preliminaries
  - phylogenies
  - phylogeny reconstruction
  - gene-order data
  - distance measures
- Inversion distance and inversion medians
- GRAPPA
- Experimental setup
- Results
- Discussion

Phylogenies

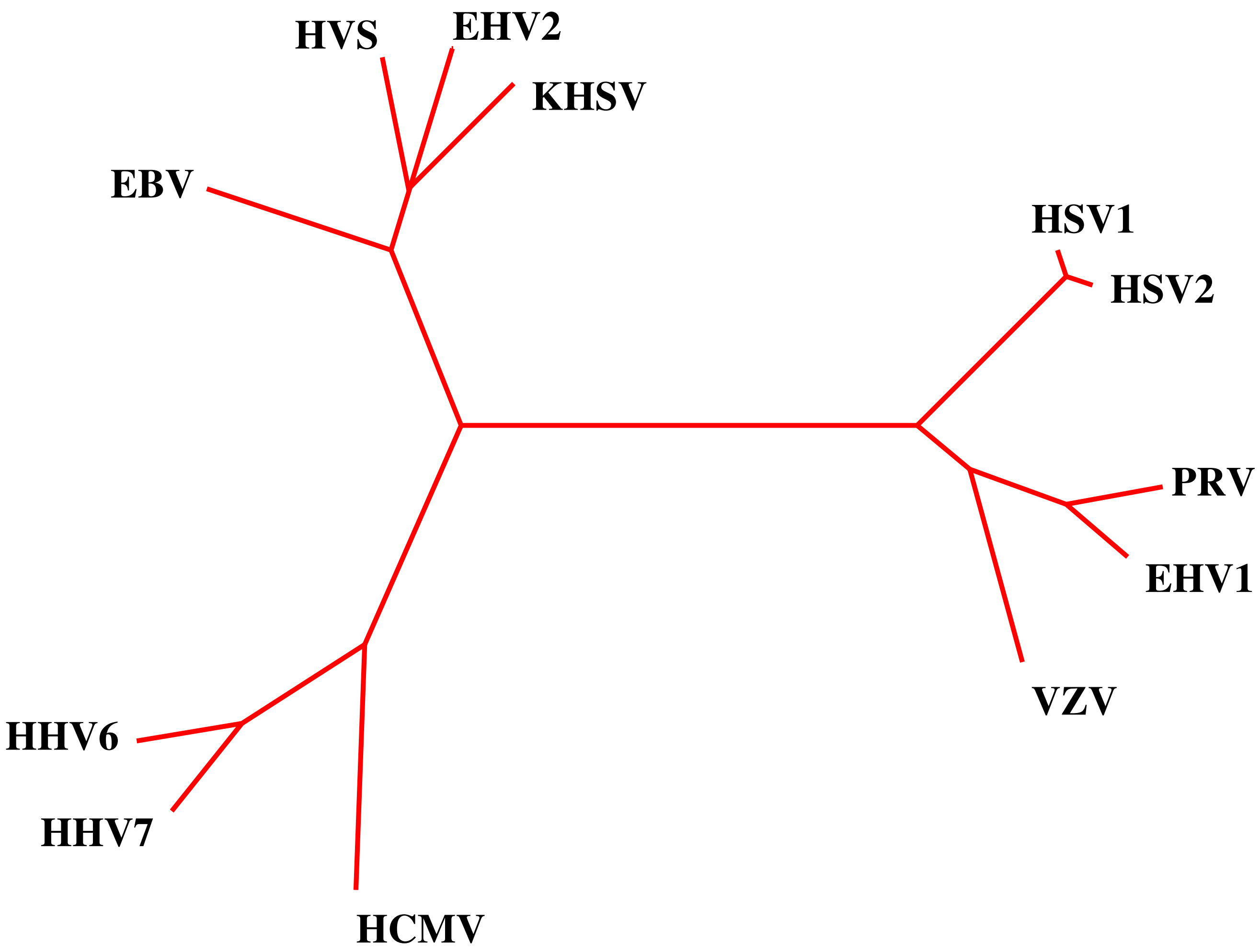
A phylogeny is a reconstruction of the evolutionary history of a collection of organisms; it usually takes the form of a tree.

- *Modern organisms are placed at the leaves and ancestral organisms occupy internal nodes.*
- *The edges of the tree denote evolutionary relationships.*

12 Species of *Campanulaceae*



Herpes Viruses that Affect Humans



Phylogeny Reconstruction

- uses molecular data (DNA, amino-acids, gene content, gene order)
- builds trees that optimize some criterion (total evolution, parsimony, maximum likelihood, etc.)
- most criteria give rise to NP-hard optimization problems
- accuracy must be tested through simulation studies (the “true tree” is not known for real datasets)

Phylogeny Reconstruction

Reconstructing phylogenies is a major component of modern research programs in many areas of biology and medicine:

- *pharmaceutical research for drug discovery*
- *understanding rapidly mutating viruses (such as HIV)*
- *designing genetically enhanced organisms*
- *explaining and predicting gene expression*
- *most centrally, understanding genomic evolution*

Gene-Order Data

- Biologists can identify specific genes and ascertain their ordering and even their orientation along a given chromosome.
- Each chromosome can thus be viewed as an ordering of signed (oriented) genes.
- When each gene appears once and all species have the same collection of n genes, the data take the form of *signed permutations* of the set $\{1, 2, \ldots, n\}$.

Gene Orders

- Biologists conjecture that genomes evolve in part through rearrangements of the gene orders.
- Such rearrangements are thought to be rare and thus give valuable information about ancient evolutionary history.
- Rearrangement mechanisms include *inversion* (well documented), *transposition* (conjectured, but not verified), and *inverted transposition* (ditto).

Genomic Distances

- Evolutionary Distance: the *historical* number of evolutionary events that separates two genomes (the sum of the number of events on the paths to their lowest common ancestor).
- Edit Distance: the *smallest* number of evolutionary events needed to transform one signed permutation into the other.
- Breakpoint Distance: the number of pairwise adjacencies present in one genome but not the other.

Genomic Distances (cont'd)

- Evolutionary distances are unknown, except in simulations.
- Breakpoint and inversion-only edit distances can be computed in linear time.

  (The latter involves nontrivial combinatorics derived by Hannenhalli and Pevzner (1995), Berman and Hannenhalli (1996), Kaplan, Shamir and Tarjan (1997), and Bader, Moret, and Yam (2001).
- Other distances (including weighted or corrected versions) can only be approximated.

Median Problem for Gene Orders

Given 3 gene orders, represented as 3 signed permutations π_1, π_2, and π_3, find a 4th permutation π_m that minimizes the sum of the distances

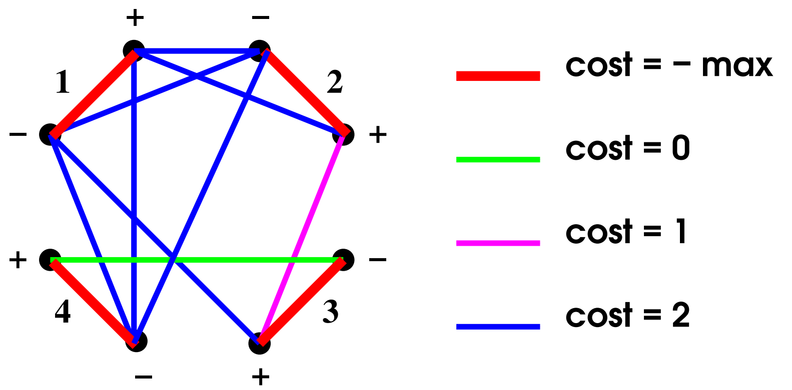
$$d(\pi_1, \pi_m) + d(\pi_1, \pi_m) + d(\pi_1, \pi_m)$$

The problem is NP-hard for breakpoint and inversion distances.

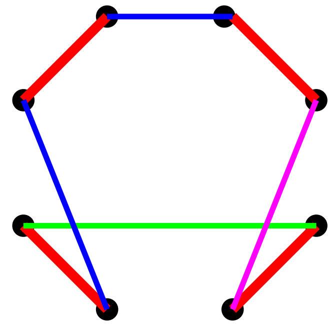
Median Problem for Breakpoints

Sankoff showed to to convert this problem to the Travelling Salesperson Problem.

+1 −2 +4 +3
+1 +2 −3 −4
+2 −3 −4 −1



edges not shown have cost = 3



an optimal solution corresponding to genome +1 +2 −3 −4

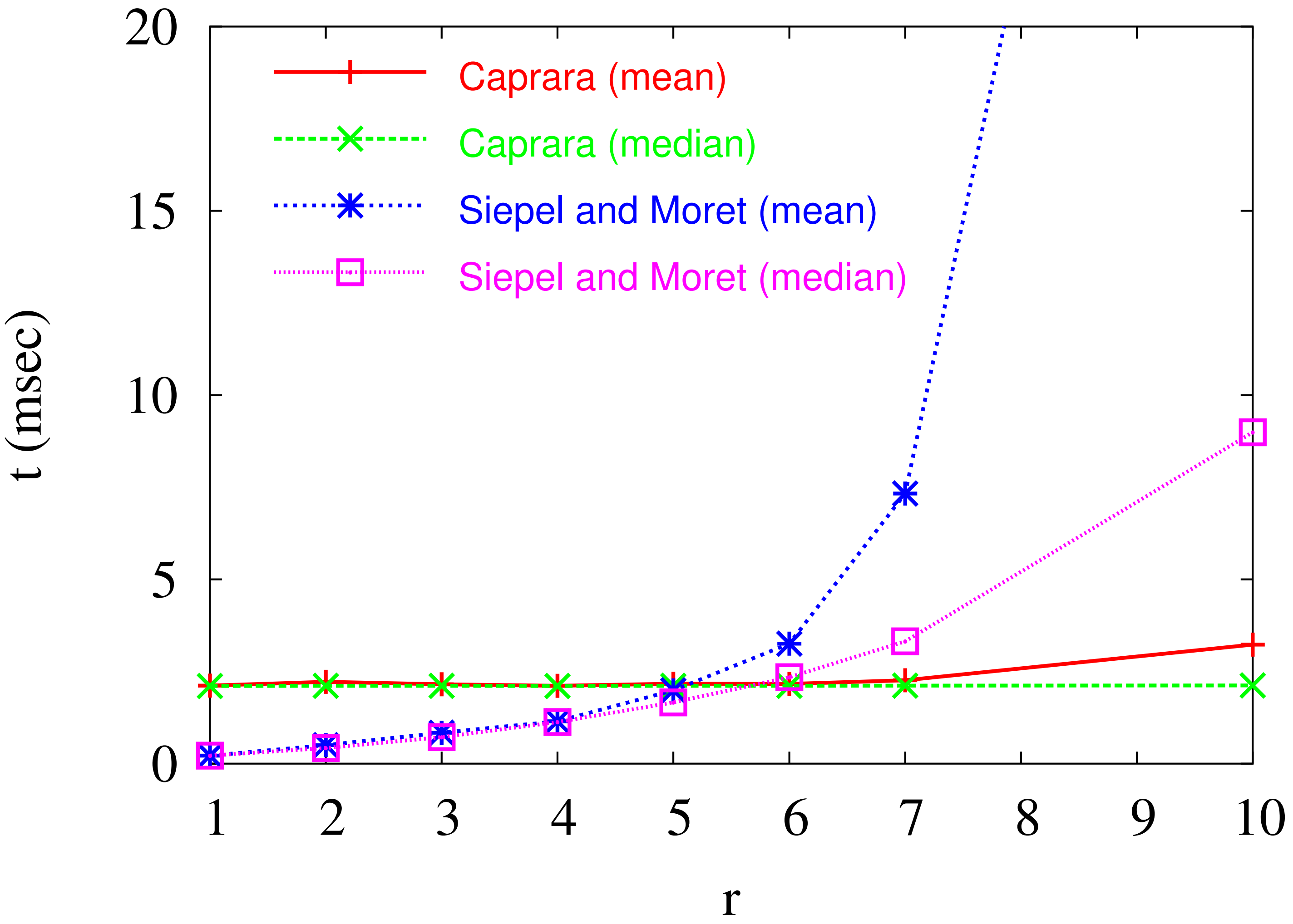
Adjacency A B becomes an edge from A to −B

The cost of an edge A −B is the number of genomes that do NOT have the adjacency A B

Median Problem for Inversions

- No known reduction to a well studied optimization problem such as TSP.
- Siepel and Moret (2001): a branch-and-bound search algorithm that performs well in practice for small distances.
- Caprara (2001): an integer programming formulation that works better for larger distances.

Inversion Medians in Practice



GRAPPA

Our software for phylogenetic reconstruction from gene-order data.

- *Inspired by Sankoff and Blanchette's* `BPAnalysis`*, but over one billion times faster.*
- *Explores all possible trees (over 13 billion trees for just 13 genomes) and scores them.*
- *For each tree, uses a Steinerization process for local improvement, iterating until convergence.*
- *Includes exact inversion and breakpoint median computations.*

www.cs.unm.edu/~moret/GRAPPA/

Inversion vs. Breakpoint Medians

- Inversion medians are much slower to compute (several orders of magnitude).
- Inversion medians are much more likely to be unique and much less likely to equal one of the three given permutations.
- Inversion medians lead to much more effective pruning (several orders of magnitude).
- Inversion medians are as fast as breakpoint medians and lead to much better results (resolution, ancestral genomes).

Real Dataset: Campanulaceae

12 Campanulaceae and 1 outgroup, 105 genes.
Low average pairwise distances:
a dataset with limited divergence.

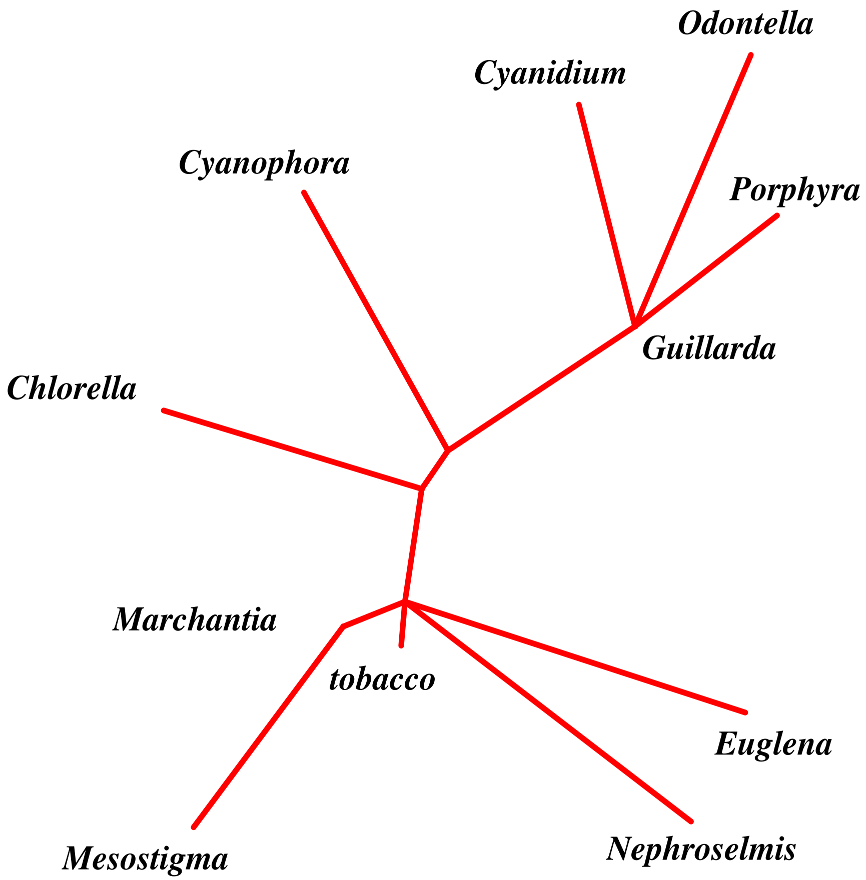
- Runs in less than one hour by pruning over 99.999% of the trees: only $100,000$ of 13.7 billion are examined. (`BPAnalysis` would have taken several centuries.)
- Returns trees with inversion score of 64—best scores achieved on this dataset; breakpoints medians yield scores of 67.

Real Dataset: 11 Green Plants

11 photosynthesizing plants, 33 genes.
High average pairwise distances:
a dataset with high divergence.

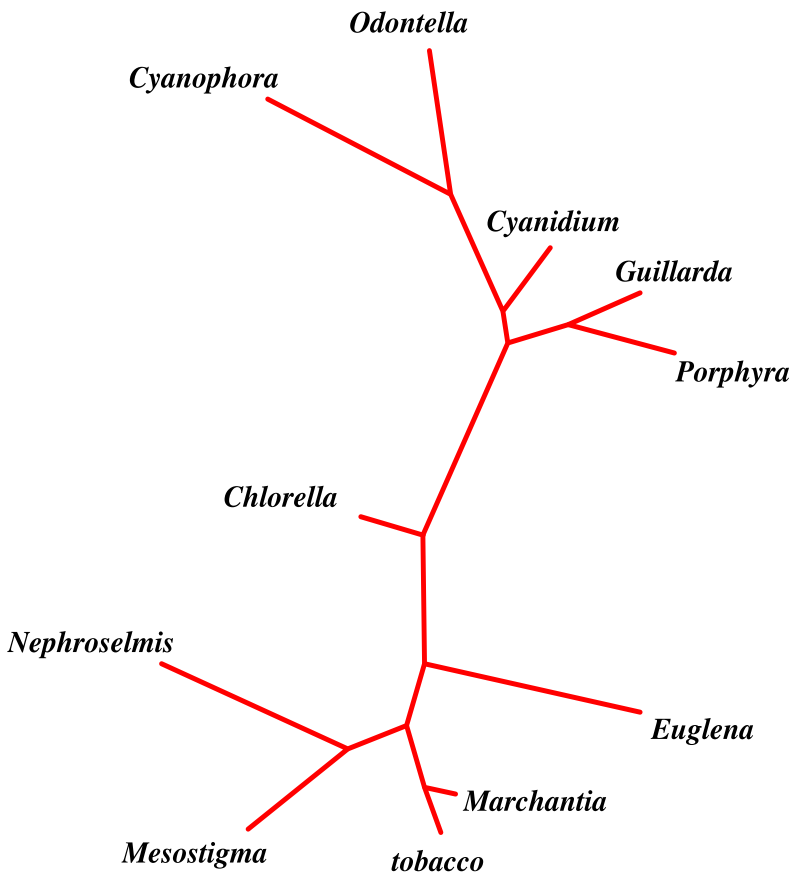
- Breakpoint medians required 14 hours of computation (pruning 25% of the trees) and returned best inversion scores of 95.
- Inversion medians required 15 mins of computation (pruning 99.9% of the trees) and returned best inversion scores of 82.

11 Green Plants: Results



breakpoint medians

4 internal edges, score 95



inversion medians

all 8 internal edges, score 82

Simulation Studies

- Methodology: generate model trees (uniform and birth-death); on each tree, evolve sequences (using `r8s` software) under various evolutionary models and parameters.
- Goals: measure and evaluate
  - speed
  - topological accuracy of reconstructed trees
  - accuracy of reconstructed ancestral genomes
  - robustness

Speed

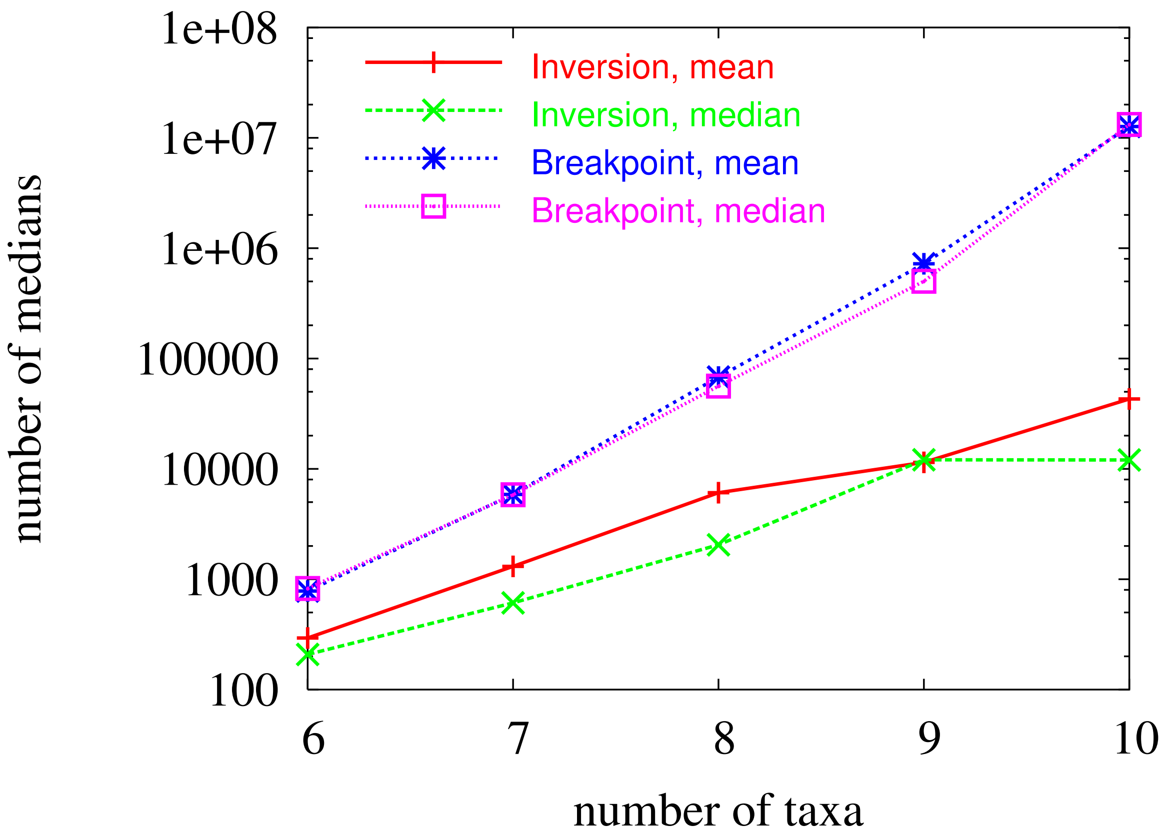
We tested on various trees of 10 and 11 taxa, with 50 and 100 genes, using various mixes of inversions, transpositions, and inverted transpositions.

Median ratios of running times for 11 taxa (breakpoint medians / inv. medians):

	$(1,0,0)$		$(\frac{2}{3},\frac{1}{6},\frac{1}{6})$		$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$	
	$r=2$	$r=4$	$r=2$	$r=4$	$r=2$	$r=4$
$n=50$	1.02	0.93	0.99	0.99	1.22	1.13
$n=100$	1.01	1.00	1.01	1.02	1.72	1.43

Pruning

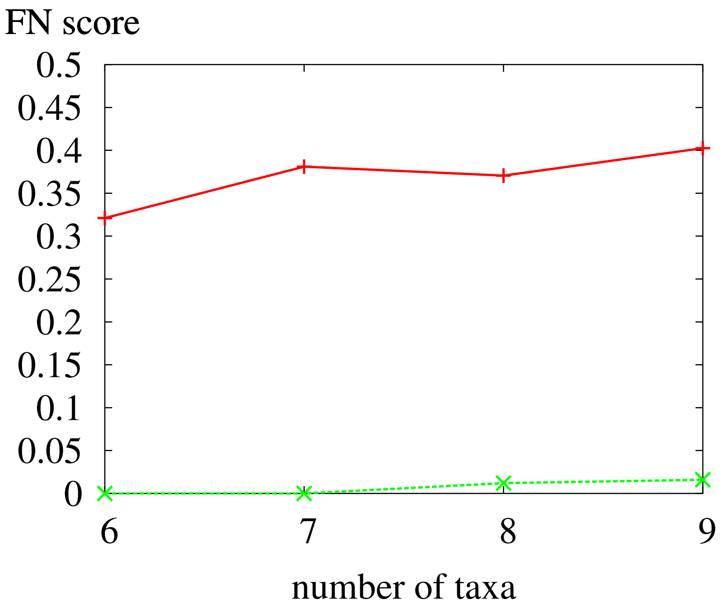
Number of calls made to median procedure, 30-gene problems, $r = 2$:



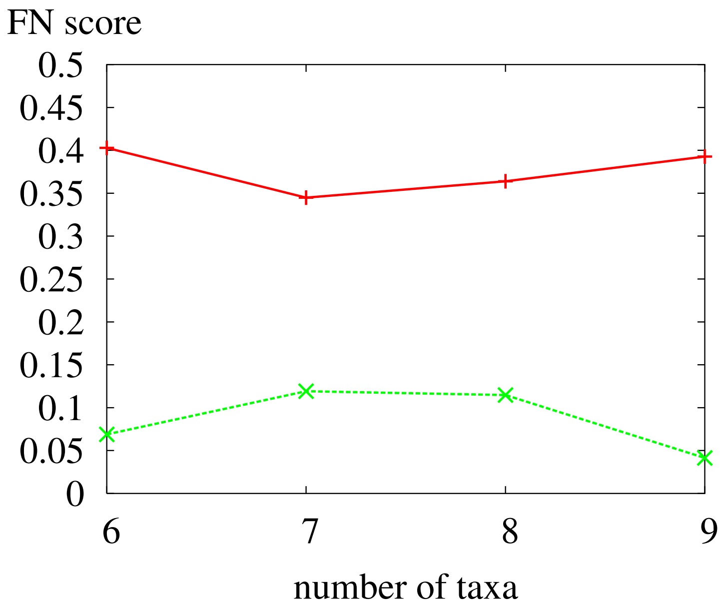
Topological Accuracy

False negative rate, 30-gene problems, $r = 2$, two ITT settings:

breakpoint medians   inversion medians



pure inversions $(1, 0, 0)$



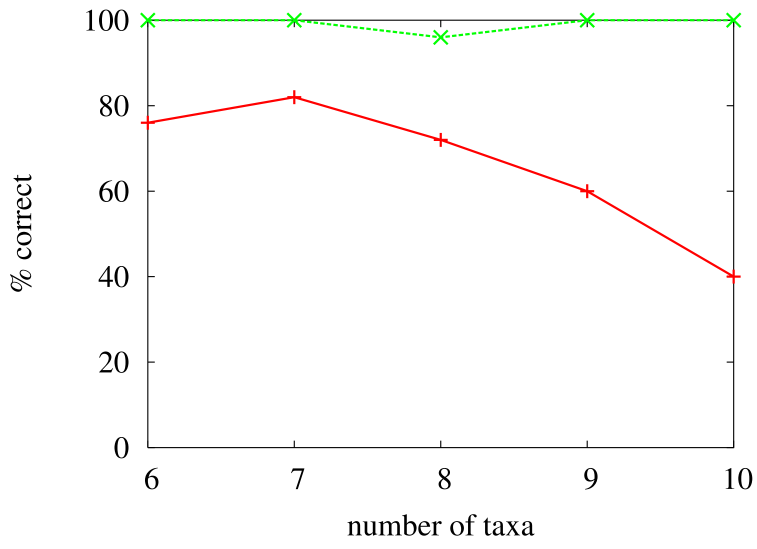
all three events $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

Topological Accuracy

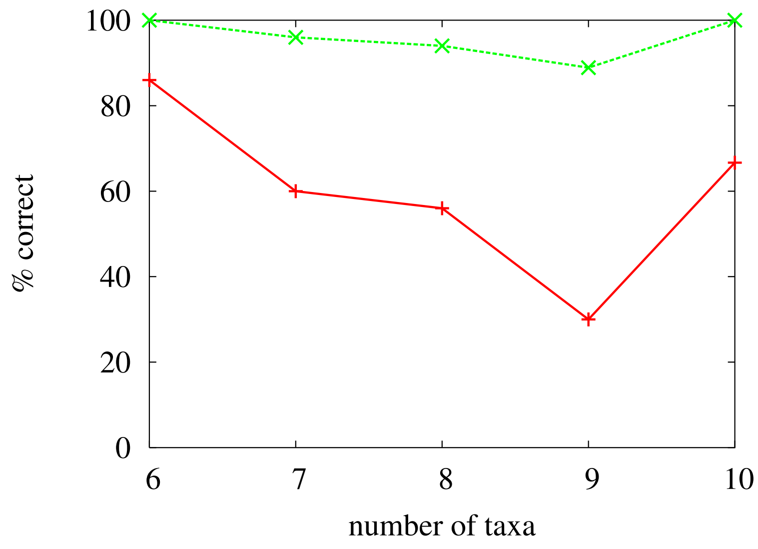
How often is the true tree among the top-scoring trees, 30-gene problems, inversion-only scenario, two rates:

breakpoint medians

inversion medians



$r = 2$



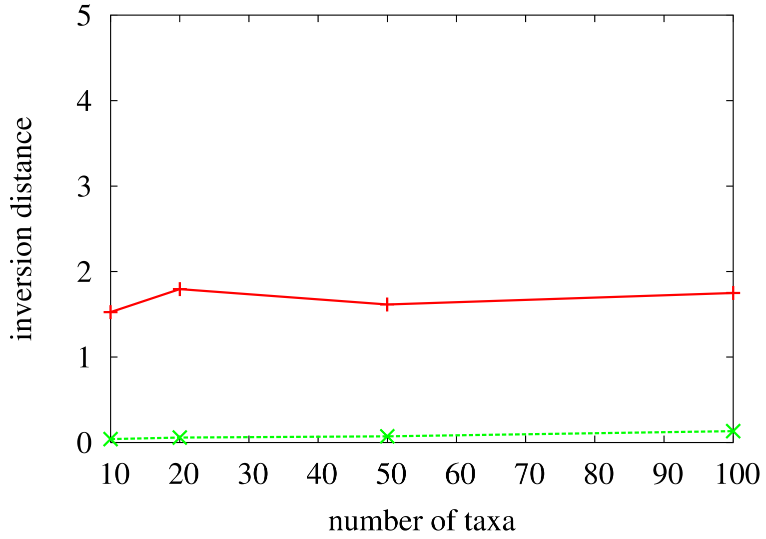
$r = 3$

Ancestral Genomes

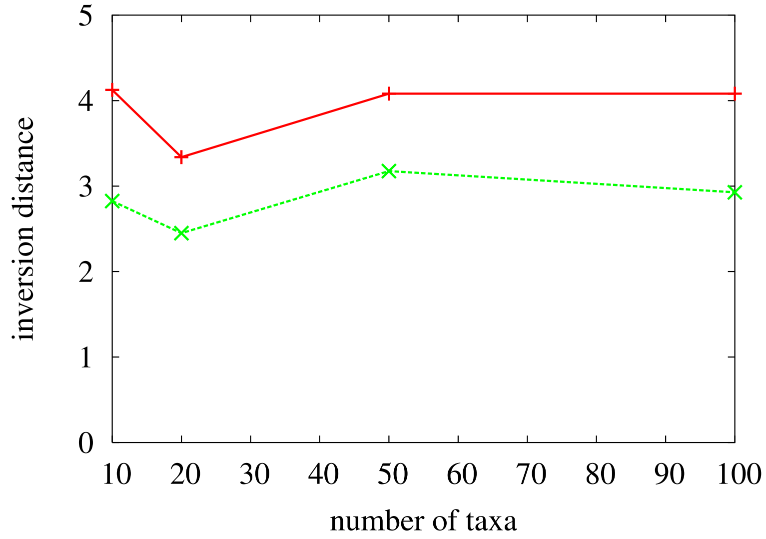
Cumulative distance from true labels, 30-gene problems, $r = 2$, best case (inversions, $(1, 0, 0)$) vs. worst-case (transpositions, $(0, 1, 0)$),

breakpoint medians

inversion medians



$(1, 0, 0)$

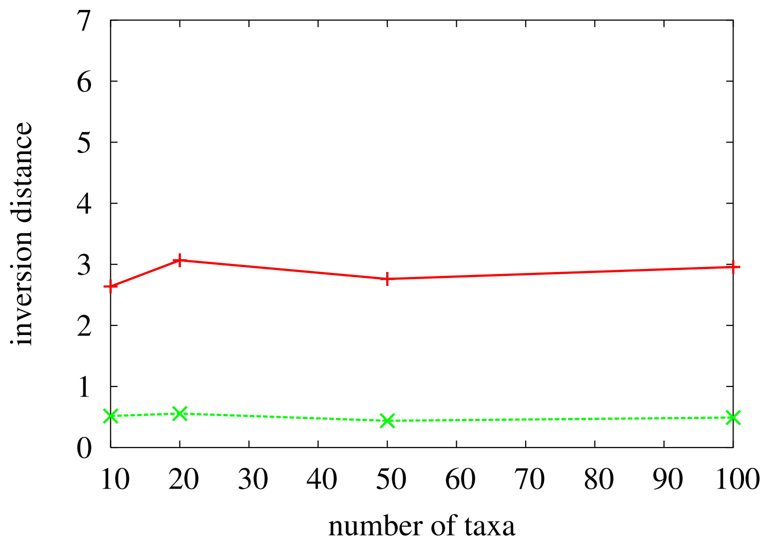


$(0, 1, 0)$

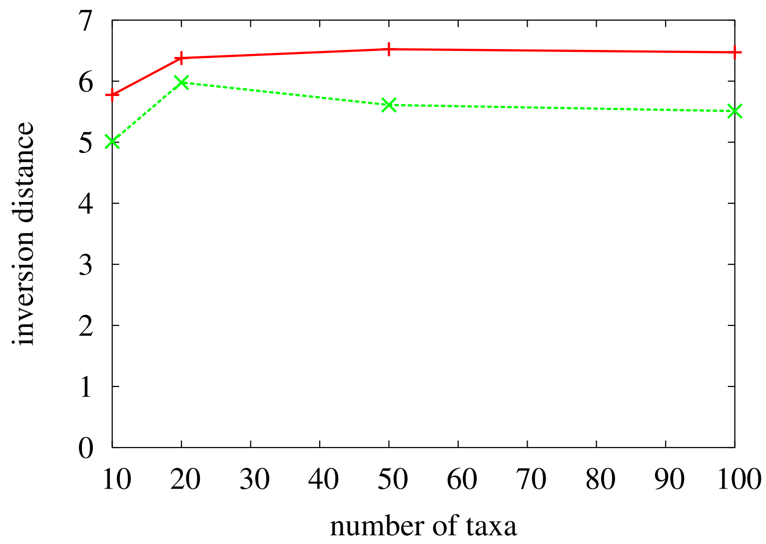
Ancestral Genomes

Cumulative distance from true labels, 30-gene problems, $r = 3$, best case (inversions, $(1,0,0)$) vs. worst-case (transpositions, $(0,1,0)$),

breakpoint medians inversion medians



$(1,0,0)$



$(0,1,0)$

Robustness

- Inversion medians suffer more from model mismatches than breakpoint medians. But:
- Inversion medians outperform (speed, resolution, ancestral genomes) breakpoint medians under all scenarios.
- A true ITT median (with adjustable weights for the various events) would be preferable, but inversion medians are the best tool to date.

Conclusions

We have demonstrated unequivocally, through a large series of experiments, that inversion medians outperform breakpoint medians

- in overall speed
- in tree resolution (false negative rate)
- in accuracy of reconstructed ancestral genomes.

These findings hold up to unrealistically large evolutionary rates and up to the largest trees we could compute with GRAPPA 1.6 (12–14 taxa for a full search, 100 taxa to label a fixed tree).